ITSC International Testing Services Center G-TELP™ General Tests of English Language Proficiency

An Examination of G-TELP Level 2 Reading Passages Over Time

2025

ITSC Research Series
Report ITSC-2025-03

An Examination of G-TELP Level 2 Reading Passages Over Time

ITSC Research Series

Report ITSC-2025-03

Prepared by Karl Sarvestani, PhD

Abstract

This study investigates diachronic changes in the linguistic quality of G-TELP Level 2 reading passages from two time periods: 2018–2020 and 2021–2024. Using automated text analysis tools, 1,333 variables across 288 passages were evaluated for readability, lexical richness, syntactic complexity, textual cohesion, and sentiment. Results revealed statistically significant improvements, particularly in lexical richness and cohesion, with genre-specific differences observed. Business Letters and Magazine Articles demonstrated the most comprehensive enhancements, while Biography texts showed notable lexical improvements, and Encyclopedia Articles improved in both lexical richness and cohesion. The over-representation of lexical variables among significant changes reflects effective targeted editorial intervention. Findings demonstrate the effectiveness of efforts to improve text quality in test materials. This study contributes foundational evidence to the under-examined area of linguistic evaluation in standardized test development and provides a basis for potential future research of textual features and broader cross-exam comparisons.

Background

For many years, International Testing Services Center (ITSC) has sought ways to strengthen the quality and consistency of its assessment materials, including the Level 2 reading passages for its General Test of English Language Proficiency (G-TELP), with many of those efforts focused on improving test-taker engagement. High-quality reading passages not only provide an accurate assessment of language proficiency but also contribute to a more positive and engaging testing experience. Therefore, to evaluate the effectiveness of ongoing in-house improvements, this study explores the extent of change over time in

quantifiable linguistic features that characterize text quality, including readability, lexical richness, syntactic complexity, textual cohesiveness, formality, and sentiment.

Text quality, in this context, extends beyond basic readability and linguistic properties to include higher-level elements of discourse. Readability, lexical richness, syntactic complexity and textual cohesiveness provide insight into how passages might influence test-taker engagement. These factors are the focus of this study both because they have been intentional targets of ITSC's enhancement efforts and because of their potential impact on the test-taker experience.

The relationship between linguistic features and reading comprehension is well documented. For instance, lexical properties of a text have been shown to significantly influence comprehension outcomes (Wright & Cervetti, 2017; Zhang & Zhang, 2020) with research suggesting that second language readers comprehend text most effectively when they recognize 95–98% of the vocabulary (Laufer & Ravenhorst-Kalovski, 2010). Moreover, some research has identified that grammatical knowledge may be as strong a predictor of reading comprehension as vocabulary knowledge (Choi & Zhang, 2018), a finding that highlights the importance of syntactic complexity to text quality assessment.

Beyond vocabulary and grammar, discourse-level features also play a crucial role in test-taker engagement. For instance, textual cohesiveness, particularly in expository genres such as those found in the G-TELP Level 2 reading passages, has been shown to support reading comprehension (Schmitz et al., 2017). In addition, sentiment and emotional tone have been found to influence test takers, as positive emotions can enhance engagement with academic tasks, including test-taking (Sinatra et al., 2015). Although students' emotional response to a topic may not directly affect learning and problem-solving, it can strongly influence their engagement by affecting their interest and motivation in an academic domain (Pekrun & Linnebrink-Garcia, 2012). Therefore, to foster a more engaging experience, G-TELP writers and editors have actively worked to create

emotionally neutral texts by minimizing potentially distressing content and avoiding topics such as violence, death, and controversial social and political issues.

The current study aims, through analysis of linguistic and discourse features, to investigate diachronic changes in G-TELP Level 2 reading passages. Specifically, it examines how these features—including lexical richness, syntactic complexity, global cohesion, and readability—differ between two distinct time periods in the test's development history. In doing so, the study also assesses the extent to which ITSC's ongoing efforts to enhance content development processes have contributed to producing reading materials of increasingly higher quality and consistency for English language proficiency assessment.

This research addresses a notable gap in the literature concerning the linguistic characterization of English proficiency test materials. In fact, while an ideal extension of the current study would involve a comparative analysis of G-TELP reading passages with those from other standardized proficiency assessments, such a comparison is currently limited by the scarcity of relevant published research. Consequently, the present study makes an important contribution by providing foundational evidence in this underexamined area and offers insight that can inform both test development practices and future research in language assessment.

Method and Materials

The materials for this project included 72 sets of reading passages drawn from retired G-TELP Level 2 tests, with 36 sets selected from each of two time periods: 2018–2020 and 2021–2024. Each set included four distinct types of texts—Biography, Magazine Article, Encyclopedia Article, and Business Letter—resulting in a total of 288 passages for analysis. Automated text analysis was conducted using a suite of freely available tools designed to generate a broad set of measures related to readability, lexical richness, text cohesion, sentiment analysis, and

syntactic complexity. The resulting measurements were then examined to develop comparative text quality profiles for passages from each of the two time periods.

The readability indices examined were measured using the Automatic Readability Tool for English (Choi & Crossley, 2022) and include traditional metrics such as the Flesch-Kincaid Grade Level (Kincaid et al., 1975), the Simple Measure of Gobbledygook (McLaughlin, 1969), and the Dale-Chall Readability Formula (Chall & Dale, 1995), alongside more modern measures such as the Coh-Metrix L2 Readability Index (Graessar et al., 2004) and the Crowdsourced Algorithm of Reading Comprehension (Crossley et al., 2019a). Lexical richness was assessed using the Tool for the Automated Analysis of Lexical Diversity (Kyle et al., 2021) as well as the Tool for the Automatic Analysis of Lexical Sophistication (Kyle et al., 2018), which generates over 500 indices encompassing both classical and modern lexical diagnostics. These include text-level coverage measures—such as the percentage of words, bigrams, and trigrams within a text captured by a given index—as well as detailed coverage information for individual lexical items and multiword sequences. These indices also include several measures of lexical frequency based on corpora such as the Corpus of Contemporary English (COCA: Davies, 2008).

Text cohesion was evaluated through the Tool for the Automatic Analysis of Cohesion (Crossley et al., 2019b), which computes more than 150 indices reflecting both local and global cohesion. This includes multiple type-token ratio measures (calculated for parts of speech, lemmas, bigrams, and trigrams), adjacent overlap indices at the sentence and paragraph levels, and various indices of connective use. Finally, syntactic complexity was measured using the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (Kyle, 2016; Lu, 2010). TAASSC provides traditional indices (e.g., mean length of T-unit) and finer-grained measures of both phrasal complexity (e.g., number of adjectives per noun phrase) and clausal complexity (e.g., number of adverbials per clause). Additionally, it includes indices informed by usage-based approaches to language acquisition, utilizing frequency profiles of verb-argument constructions to capture nuanced aspects of syntactic development.

Table 1 below summarizes the categories and numbers of measures examined and includes an example of a relevant analysis for each category.

Table 1
Measures of Linguistic Characteristics

Category	Relevant Analysis	Number of indices
Readability	Flesch Reading Ease	8
Lexical Diversity	Type-token ration	37
Lexical Sophistication	Bigram coverage	522
Syntactic Sophistication	Sentence length	353
Cohesion	Word2Vec	167
Sentiment Analysis	Negative emotion words	246
Total		1,333

Due to the abundance of indices of lexical and syntactic sophistication, it was expected that these categories would be the most likely, statistically, to show significant differences between the two time periods under examination, even if only by pure chance. In other words, since 42% of the variables examined were indicators of lexical richness, there would likely be a proportionate number of the variables with significant differences between the early and late text passages. However, if they comprised, for example, a larger portion of the observed differences, then that would suggest that there had been a larger change in lexical richness than there had been in other aspects of the text. With this consideration in mind, any significant differences observed were examined with regard to the proportion of total indices of the category measured.

Of the original 1,333 variables examined, 19 were excluded from analysis and therefore removed from discussion below because they were constant throughout the population, meaning that they could not contribute to any changes observed. For each of the remaining 1,314 variables, two-tailed, two-sample *t*-tests were performed. The type of *t*-test was determined algorithmically based on the results of Levene's test for equality of variances. Where the variances were unequal (*p*-value from Levene's test < .05), the function used two-

tailed, two-sample Welch's *t*-tests. Conversely, where the variances were equal (*p*-value from Levene's test ≥ .05), the function used the standard two-tailed, two-sample *t*-test (assuming equal variances). Due to the large number of tests run and attendant elevated chance of Type I error, the Benjamini-Hochberg FDR procedure was used.

Results

Changes observed between the early (2018–2020) and late (2021–2024) reading passages are summarized in Table 2, which presents means, standard deviations, *t*-statistics, and the category of each measure for the 16 variables that were found to have significant differences ($p < .05$) between the two time periods. At this time, the remaining 1,298 variables, for which no significant change was found, were excluded from analysis.

Table 2

Changes Observed Between Early and Late Texts

Features	Category	*t*	Early Mean (SD)	Late Mean (SD)
New.Dale.Chall.Readability.Formula	Readability	-3.85	9.62 (0.77)	9.25 (0.84)
nn_all_nominal_deps_struct	Syntactic	-3.48	0.17 (0.07)	0.14 (0.07)
COCA_fiction_bi_prop_70k	Lexical	3.48	0.40 (0.05)	0.42 (0.06)
COCA_fiction_bi_prop_80k	Lexical	3.75	0.41 (0.05)	0.44(0.06)
COCA_fiction_bi_prop_90k	Lexical	3.77	0.42 (0.05)	0.45(0.06)
COCA_fiction_bi_prop_100k	Lexical	3.74	0.44 (0.06)	0.46(0.06)
basic_ntokens	Lexical	3.51	311 (38.19)	325 (27.89)
basic_ntypes	Lexical	4.56	166 (18.44)	175 (15.27)
basic_nfunction_tokens	Lexical	5.07	153 (17.58)	163 (12.81)
basic_nfunction_types	Lexical	4.46	46.92 (5.42)	49.97 (6.16)
root_ttr_aw	Lexical	4.14	9.40 (0.07)	9.70 (0.57)
root_ttr_cw	Lexical	3.73	9.47 (0.079)	9.80 (0.74)
adjacent_overlap_2_fw_para	Cohesion	4.89	0.56 (0.09)	0.60 (0.06)
addition	Cohesion	-3.93	0.03 (0.01)	0.03 (0.01)
negative_logical	Cohesion	3.57	0.005 (0.004)	0.01 (0.004)
Indadj_GI	Sentiment	3.92	0.002 (0.003)	0.004 (0.005)

For each of the language features examined—readability, lexical richness, text cohesion and sentiment—at least one index displayed a significant difference. Notably, 10 of the 16 features shown in Table 2 are directly related to lexical richness. Lexical features constituted the largest single category of variables included in the analysis, accounting for 42% of the total features evaluated. However, they represented 63% of the features that displayed diachronic change.

While not quite as pronounced, cohesion features were also more prominent than might have been expected. Although only 13% of the total features evaluated measured textual cohesion, such features accounted for 19% of features with observed statistical differences between the earlier and later texts.

Overall, relevant variables included four COCA features that were the proportion of bigrams in the text from among the 70,000–100,000 most frequent bigrams in COCA fiction corpus. The features "basic_ntokens" and "basic_ntypes" are the simple counts of word tokens and types, respectively. Similarly, "basic_nfunction_tokens" and "basic_nfunction_types" are the simple counts of function word tokens and function word types. Also included was a slightly more complex measure, "root_ttr_aw", which is the number of word types for all words, divided by the square root of the number of all word tokens, also known as "Guiraud's Index". The feature "root_ttr_cw", is the same calculation, but counts only content words.

Of the remaining six features, three are directly related to textual cohesiveness. The feature "adjacent_overlap_2_fw_para" is a measure of adjacent, two-paragraph overlap lemmas in function words. "Addition" is the number of addition words (e.g., "and," "also") in each text. The feature "negative_logical" refers to the number of negative logical connectives (e.g., "admittedly," "alternatively," "although") in the texts. The remaining three features are "X.New.Dale.Chall.Readability.Formula," which is the New Dale-Chall readability index, "Indadj_GI," which counts instances of adjectives from a set of 637 adjectives used to describe people (e.g., "abominable," "absent," "absent-minded," "accountable") and, finally, "nn_all_nominal_deps_NN_struct," which is the count of nouns used as nominal dependents per nominal, excluding pronouns. The general contributions of these different classes of

measures related to text quality are illustrated in Figure 1 below, which demonstrates that as classes, syntactic complexity and lexical richness make the largest contributions to the differences between the older and newer texts.

Figure 1

Mean Effect Size by Feature Category

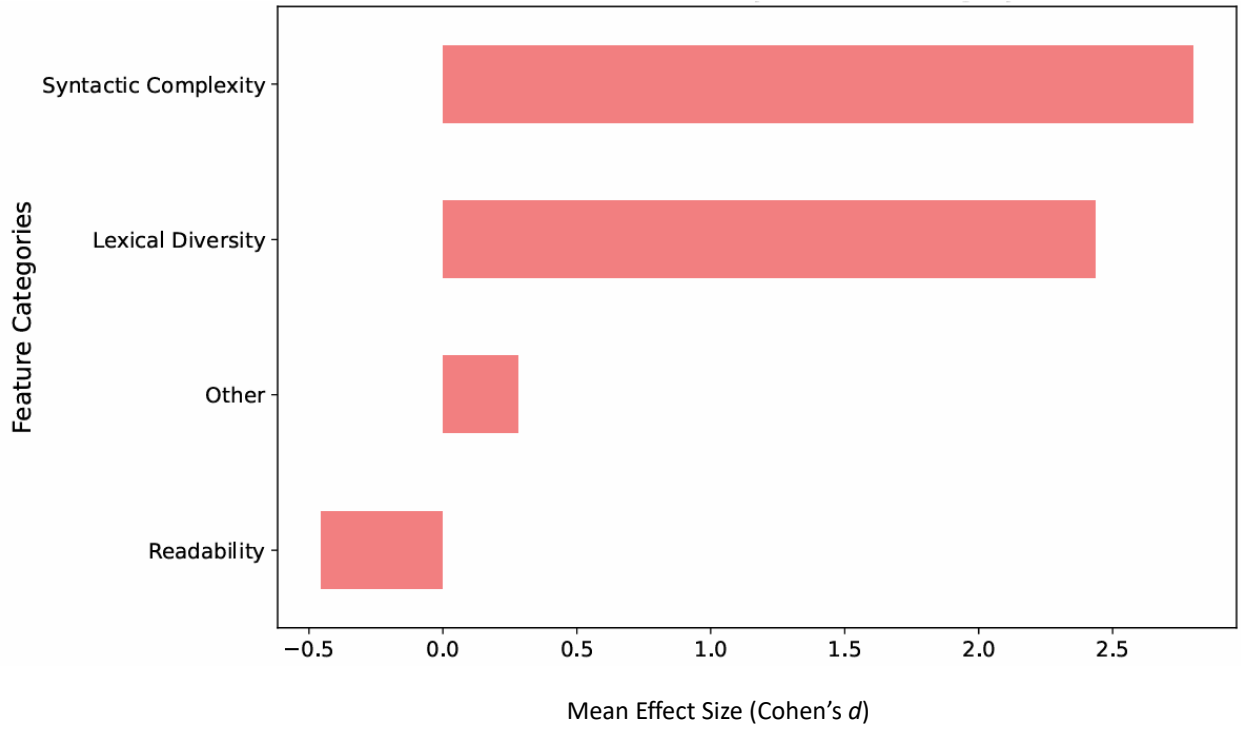

Figure 2 illustrates the effect sizes, in isolation, of each of the variables for which a significant difference was found between the earlier and later texts. Within second language research, a Cohen's *d* value of 0.4–0.69 is considered a "small" effect size, 0.7–0.99 a "medium" effect size, and 1.0 or greater a "large" effect size (Plonsky & Oswald, 2014). From the figure, it is clear that the basic counts of word types and tokens ("basic_ntokens", "basic_ntypes") as well as the basic count of function word tokens ("basic_nfunction_tokens"), the lexical overlap in function words between adjacent paragraphs ("adjacent_overlap_2_fw_para"), and the sentiment index ("Indadj_GI"), have the largest effects. Their positive directionality indicates an increase in the

variable measured. Both the basic count of function word types (“basic_nfunction_types”) and the Guiraud’s Index of all words (“root_ttr_aw”) have medium effects in the positive direction. The remaining effects are all small in size, with the number of “addition” words, the Dale-Chall readability formula (“X.New.Dale.Chall.Readability.Formula”), and the per-nominal count of nouns used as nominal dependents, (“nn_all_nominal_deps_struct”) being in the negative direction, indicating a decrease in these indices.

Figure 2

Effect Sizes for Individual Features

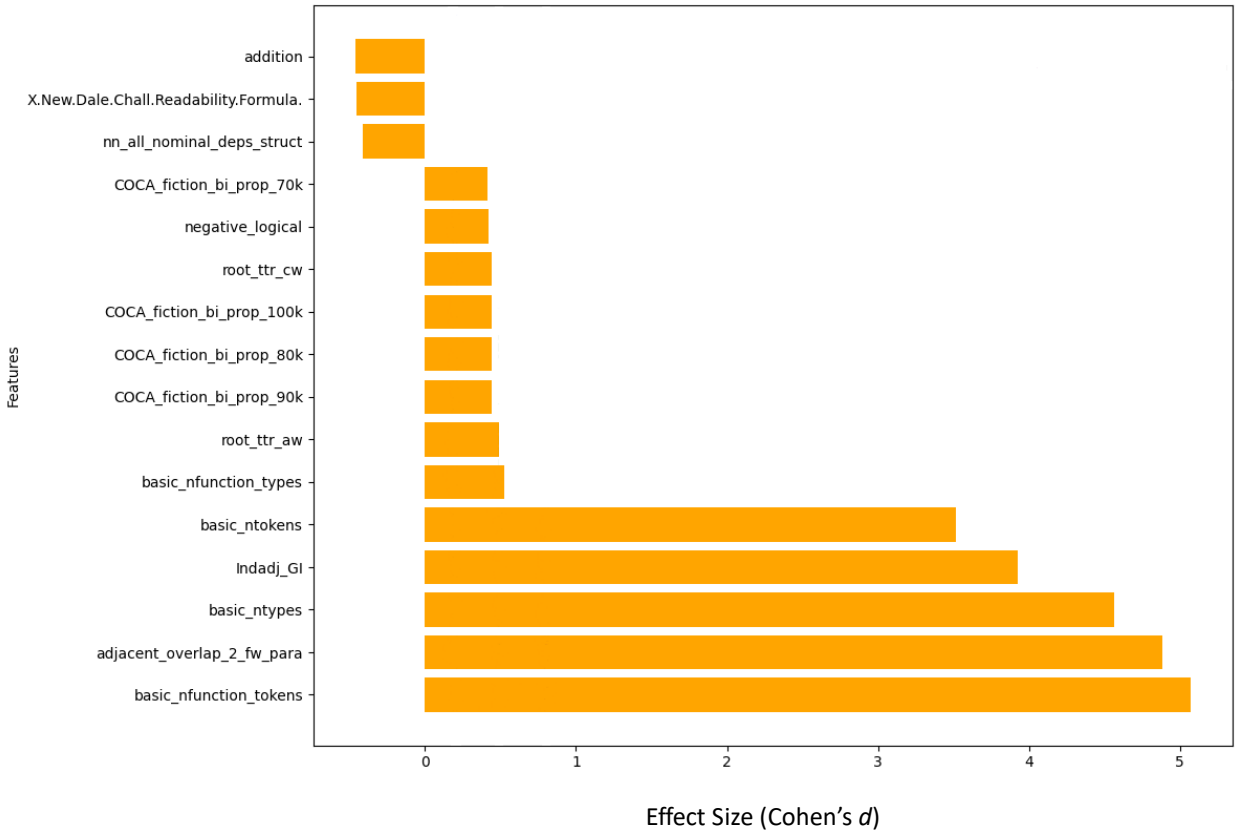

Table 3 shows the means and standard deviations by genre for each variable that was found to have a significant difference between the early- and late-range Biography texts. Out of the seven variables for which there were significant differences between the early range and late

range Biography texts, all seven were measures of lexical richness.

Table 3

Means and Standard Deviations for Earlier vs. Later Biographies

Variable	Category	Early Mean (SD)	Late Mean (SD)
COCA_fiction_bi_prop_70k	Lexicon	0.39 (0.03)	0.41 (0.04)
COCA_fiction_bi_prop_80k	Lexicon	0.40 (0.03)	0.42 (0.04)
COCA_fiction_bi_prop_90k	Lexicon	0.41 (0.03)	0.44 (0.04)
COCA_fiction_bi_prop_100k	Lexicon	0.43 (0.03)	0.45 (0.04)
basic_ntypes	Lexicon	178 (11.26)	185.72 (10.19)
root_ttr_aw	Lexicon	9.65 (0.52)	9.96 (0.51)
root_ttr_cw	Lexicon	10.05 (0.56)	10.46 (0.55)

Table 4 presents the same information for the variables that were statistically significant for the Encyclopedia Articles.

Table 4

Means and Standard Deviations for Earlier vs. Later Encyclopedia Articles

Variable	Category	Early Mean (SD)	Late Mean (SD)
basic_ntokens	Lexicon	323.81 (15.40)	331.17 (11.71)
basic_ntypes	Lexicon	172.64 (11.75)	178.47 (9.73)
root_ttr_cw	Lexicon	9.61 (0.71)	9.94 (0.61)
adjacent_overlap_2_fw_para	Cohesion	0.56 (0.07)	0.61 (0.07)

The only overlap between the two genres (Biography and Encyclopedia Article) occurs with "basic_ntypes" and "root_ttr_cw," indicating that both genres have improved in their lexical richness. However, while the inclusion of low-frequency vocabulary has significantly increased in the Biography texts, as illustrated by the several COCA variables, such a change is not readily apparent in the Encyclopedia Articles. Rather, while the only significant changes observed in the Biography texts were in the category of lexical richness, the Encyclopedia Articles showed improvement in the textual cohesiveness feature "adjacent_overlap_2_fw_para".

Results of the Business Letter analyses are shown in Table 5. These illustrate not only total overlap with the significant changes in both the Biography text and Encyclopedia Articles, but also changes over time across nearly every variable, including measures of lexical richness, syntactic complexity, textual cohesiveness, and readability. This suggests that the Business Letters have undergone a more complete transformation over time than either the Biography texts or Encyclopedia Articles. Notably, the Business Letters also show significant increases in both the raw length of the texts, as measured by “basic_ntokens", and the type-token ratios of all words and content words.

Table 5

Means and Standard Deviations for Earlier vs. Later Business Letters

Variable	Category	Early Mean (SD)	Late Mean (SD)
X.New.Dale.Chall.Readability.Formula.	Readability	9.35 (0.83)	8.51 (0.80)
nn_all_nominal_deps_struct	Syntax	0.17 (0.08)	0.13 (0.05)
COCA_fiction_bi_prop_70k	Lexicon	0.44 (0.06)	0.49 (0.07)
COCA_fiction_bi_prop_80k	Lexicon	0.45 (0.06)	0.51 (0.07)
COCA_fiction_bi_prop_90k	Lexicon	0.46 (0.06)	0.52 (0.07)
COCA_fiction_bi_prop_100k	Lexicon	0.47 (0.06)	0.53 (0.07)
basic_ntokens	Lexicon	250.67 (13.06)	285.08 (19.82)
basic_ntypes	Lexicon	141.31 (9.88)	157.81 (13.85)
basic_nfunction_tokens	Lexicon	130.33 (9.94)	152.69 (13.29)
basic_nfunction_types	Lexicon	43.17 (5.10)	50.31 (6.44)
root_ttr_aw	Lexicon	8.93 (0.55)	9.34 (0.58)
root_ttr_cw	Lexicon	8.94 (0.67)	9.34 (0.60)
adjacent_overlap_2_fw_para	Cohesion	0.48 (0.12)	0.59 (0.07)
negative_logical	Cohesion	0.003 (0.005)	0.01 (0.005)

Finally, Table 6 displays the early and late means and standard deviations for the twelve variables with significant differences in the Magazine Articles. Based on these results, it appears that both Business Letters and Magazine Articles have changed in substantially similar ways encompassing their lexical richness, syntactic complexity, textual cohesiveness and readability.

Table 6

Means and Standard Deviations for Earlier vs. Later Magazine Articles

Variable	Category	Early Mean (SD)	Late Mean (SD)
X.New.Dale.Chall.Readability.Formula.	Readability	10.07 (0.67)	9.33 (0.72)
nn_all_nominal_deps_struct	Lexicon	0.16 (0.06)	0.12 (0.07)
COCA_fiction_bi_prop_70k	Lexicon	0.38 (0.05)	0.41 (0.05)
COCA_fiction_bi_prop_80k	Lexicon	0.39 (0.05)	0.42 (0.05)
COCA_fiction_bi_prop_90k	Lexicon	0.40 (0.05)	0.43 (0.05)
COCA_fiction_bi_prop_100k	Lexicon	0.41 (0.05)	0.44 (0.05)
basic_ntokens	Lexicon	327.42 (13.83)	335.19 (14.45)
basic_ntypes	Lexicon	170.53 (13.23)	177.31 (11.03)
basic_nfunction_tokens	Lexicon	155.72 (11.33)	163.58 (10.35)
basic_nfunction_types	Lexicon	49.06 (4.70)	53.28 (5.40)
addition	Cohesion	0.03 (0.01)	0.02 (0.01)
negative_logical	Cohesion	0.01 (0.00)	0.01 (0.00)

Discussion

Overall, significant improvements in the G-TELP Level 2 reading passages' text quality were observed across multiple linguistic features, with notable variation across text genres between earlier and later texts. These findings provide insight into the development of text quality over time and illustrate how changes in linguistic measures contribute to a more sophisticated and engaging text.

Results indicate substantial improvements in lexical richness within every genre examined, particularly in terms of the range of vocabulary used in both content words and function words. Although measures of lexical richness were the largest category of variables examined, they were still overrepresented among the diachronic changes observed. This disproportionate occurrence is consistent with the notion that improvements in text quality are accompanied by increases in lexical richness and sophistication.

The COCA-based bigram measures (COCA_fiction_bi_prop_70k to COCA_fiction_bi_prop_100k) revealed that the more recent passages, particularly the Biography texts, Business Letters, and Magazine Articles, exhibited a higher proportion of

low-frequency bigrams, compared to their earlier counterparts. This suggests an increased use of less-common lexical items. The basic measures of word type counts also showed significant improvements across the genres. The increase in "basic_ntypes" (total word types) reflects broader use of vocabulary, as evidenced by the higher values in late texts across all genres. This trend was particularly marked in the Business Letters, the genre that initially showed the lowest total word types at 141 but experienced the largest increase of any genre, reaching 158 word types in the later texts. The increase in the root TTR for all words and content words further highlights the enhanced lexical richness due to the nonlinear relationship between types and tokens, as longer texts tend to have lower type-token ratios (Kyle et al., 2021). The fact that the ratios increased as the length of the Business Letters also increased is evidence of a more robust improvement in lexical richness than might be immediately apparent. The improved lexical richness, particularly in the use of less frequent bigrams and content words, is likely a marker of higher quality and more engaging texts (Kim et al., 2017).

There were also meaningful improvements in syntactic complexity, especially in the increased use of nominal dependents (“nn_all_nominal_deps_struct”) and the higher counts of function words (e.g., “basic_nfunction_tokens” and “basic_nfunction_types”). This trend is particularly evident in the Business Letters and Magazine Articles, which displayed significant positive changes in both the number of function words and the diversity of function words used. Function words play a crucial role in syntactic complexity by structuring sentence-level relationships, and their increased use suggests more sophisticated syntactic constructions (Field, 2008). In contrast, the nominal dependency measure ("nn_all_nominal_deps_struct") saw a slight decrease in some genres, such as Business Letters and Magazine Articles, although the effect size was small. While this suggests slightly less complex syntactic structures in certain contexts, the effect of increased function word types and tokens is far larger, and the later Business Letters and Magazine Articles seem to be more syntactically complex overall than the earlier texts. This improvement in syntactic

complexity aligns with prior research (Taguchi et al., 2013) and is consistent with the finding that the later (more recent) reading passages are generally higher-quality text.

The measures of textual cohesiveness, particularly "adjacent_overlap_2_fw_para," revealed a positive shift across Encyclopedia Articles and Business Letters. This feature, which quantifies the overlap of function words between adjacent paragraphs, indicates stronger textual cohesion in the later texts, which suggests an improvement in connections between paragraphs, making texts more coherent and easier to follow. Cohesiveness is a critical aspect of engaging writing, as it allows for smoother transitions and a more organized flow of ideas (Crossley et al., 2014).

The analysis of readability, as measured by the New Dale-Chall Readability Formula ("X.New.Dale.Chall.Readability.Formula"), showed a slight decrease in readability scores across Business Letters and Magazine articles. This decrease suggests that while the texts have become more lexically and syntactically sophisticated, they may also have become slightly easier to read. However, the relatively small effect size, measured by Cohen's *d* and observed in the readability measure, suggests that the observed changes in complexity were not drastic. While published data on the readability scores of English language proficiency exams is limited, Sibeko and van Zaanen (2021) report a Dale-Chall readability score of 9.3 for the South African Department of Basic Education grade 12 test of English as a First Additional Language, which is comparable to the Dale-Chall score of the later G-TELP Level 2 reading passage texts examined here.

The Biography texts showed the most substantial improvements in lexical richness, particularly through the increased use of low-frequency vocabulary (as evidenced by COCA bigram measures), while the Encyclopedia Articles demonstrated improvements in textual cohesiveness, but fewer significant changes in lexical richness. This could be related to relative narrativity of the two genres: Biography texts tend to be organized as a narrative of the life of their subject, and highly narrative genres tend to display lower rates of textual cohesion (Graesser et al., 20004). These contrasting trends reflect the different stylistic and

functional demands of these genres, with Biography texts allowing for greater lexical variation, and Encyclopedia Articles potentially prioritizing clarity and factual precision. The Business Letters and Magazine Articles, on the other hand, showed improvements across nearly all categories, suggesting that these genres may have been particularly in need of improvement. The substantial overlap in significant changes between these two genres may further suggest that they share linguistic similarity, in part because these are the two G-TELP Level 2 reading texts of relatively less formal genres. Informal text tends, among other things, to have lower cohesion than formal text. This is because formal text involves more preplanning than informal text and has a need to be more "precise, coherent, articulate, and convincing to an educated audience" (Graesser et al., 2014, p. 218) than less formal text. In other words, a general trend toward more formal, engaging language in the later texts would be most likely to impact genres such as Business Letters, considering the effect that formal language has on the inclusion of cohesive devices (Li & Graesser, 2017).

Conclusion

This study revealed significant advancements in the quality and consistency of G-TELP Level 2 reading passages between two time periods (2018–2020, 2021–2024) in terms of lexical richness, syntactic complexity, textual cohesiveness, and readability. Results show that ITSC's efforts to improve test materials have led to more engaging and linguistically sophisticated texts in several aspects, which is likely to enhance the overall test-taker experience. The most notable improvements were observed in lexical richness, with later texts featuring a broader range of vocabulary. These more recent texts displayed a marked increase in the use of low-frequency bigrams and a higher overall word type count.

While differences in improvements across genres were noted, with Biography texts showing the most significant lexical advances and Encyclopedia Article texts showing more improvements in cohesiveness, the overall trend points to the effectiveness of ITSC's efforts. Notably, the Business Letter and Magazine Article genres have shown substantial improvements across all measured categories.

Increases in lexical richness, syntactic complexity, textual cohesion, and readability between the two examined periods suggest that modifications to the test development process have, in fact, contributed to producing higher-quality and more consistent reading materials. These findings have implications for the ongoing development of English proficiency assessments, with potential benefits for both the accuracy of language ability measurement and the quality of test materials provided to learners.

Future research might extend this work by examining the relationship between text quality improvements and test-taker performance data to explore how changes in passage characteristics influence comprehension and item difficulty. Additionally, comparative analyses across different English proficiency exams could offer further insight into best practices for developing reading materials that align with contemporary standards in language assessment and applied linguistics.

References

Chall, J. S., & Dale, E. (1995). *Readability revisited, the new Dale-Chall readability formula.* Brookline.

Choi, J. S., & Crossley, S. A. (2022). Advances in readability research: A new readability Web app for English. In 2022 *International Conference on Advanced Learning Technologies* (ICALT) 1–5. IEEE. https://doi.org/10.1109/icalt55010.2022.00007

Choi, Y., & Zhang, D. (2018). The relative role of vocabulary and grammatical knowledge in L2 reading comprehension: A systematic review of literature. *International Review of Applied Linguistics in Language Teaching*, *59*(1), 1–30. https://doi.org/10.1515/iral-2017-0033

Crossley, S. A., Kyle, K., & Dascalu, M. (2019b). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*, 14–27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2014). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing 32,* 1–16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social order analysis. *Behavior Research Methods 49*(3), 803–821. doi:10.3758/s13428-016-0743-z

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019a). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, *42*(3–4), 541–561. https://doi:10.1111/1467-9817.12283

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Field, J. (2008). Bricks or mortar: Which parts of the input does a second language listener rely on?. *TESOL Quarterly*, *42*(3), 411–432. https://doi.org/10.1002/j.1545-7249.2008.tb00139.x

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal 115*(2), 210–229. https://doi.org/10.1086/678293

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/bf03195564

Kim, M., Crossley, S. A., & Kyle, K. (2017). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development and writing quality. *The Modern Language Journal 102*(1), 120–141. https://doi.org/10.1111/modl.12447

Kincaid, J. P., Fishburne, R. P. Jr., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. *Institute for Simulation and Training, 56.* https://doi.org/10.21236/ada006655

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral Dissertation: Georgia State University]. https://doi.org/10.57709/8501051

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly 18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

Li, H., & Graesser, A. (2017). Impact of pedagogical agents' conversational formality on learning and engagement. In *Artificial Intelligence in Education: 18th International Conference, AIED 2017, Wuhan, China, June 28–July 1, 2017, Proceedings 18,* 188-200. Springer International Publishing. https://doi.org/10.1007/978-3-319-61425-0_16

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, *12*(8), 639–646.

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In *Handbook of research on student engagement,* 259–282. Springer. https://doi.org/10.1007/978-3-031-07853-8_6

Plonsky, L., & Oswald, F. L. (2014). How big Is "big"? Interpreting effect sizes in L2 research. *Language Learning 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Schmitz, A., Gräsel, C., & Rothstein, B. (2017). Students' genre expectations and the effects of text cohesion on reading comprehension. *Reading and Writing*, *30*, 1115–1135. https://doi.org/10.1007/s11145-016-9714-0

Sibeko, J., & van Zaanen, M. (2021). An analysis of readability metrics on English exam texts. *Journal of the Digital Humanities Association of Southern Africa 3*(1), 1–11. https://doi.org/10.55492/dhasa.v3i01.3864

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist, 50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, *47*(2), 420–430. https://doi.org/10.1002/tesq.91

Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*, *52*(2), 203–226. https://doi.org/10.1002/rrq.163

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research, 26* (4)*,* 696–725. https://doi.org/10.1177/1362168820913998